Ecological Monographs, 81(3), 2011, pp. 511–525

Generalized linear mixed models for phylogenetic analyses of community structure

Anthony R. Ives[1,3] and Matthew R. Helmus[2]

[1] *Department of Zoology, University of Wisconsin, 459 Birge Hall, Madison, Wisconsin 53706 USA*
[2] *Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Kunming, Yunnan 650223 China*

Abstract. There is growing appreciation that ecological communities are phylogenetically structured, with phylogenetically closely related species either more or less likely to co-occur at the same site. Here, we present phylogenetic generalized linear mixed models (PGLMMs) that can statistically test a wide variety of phylogenetic patterns in community structure. In contrast to most current statistical approaches that rely on community metrics and randomization tests, PGLMMs are model-based statistics that fit observed presence/absence data to underlying hypotheses about the distributions of species among communities. We built four PGLMMs to address (1) phylogenetic patterns in community composition, (2) phylogenetic variation in species sensitivities to environmental gradients among communities, (3) phylogenetic repulsion in which closely related species are less likely to co-occur, and (4) trait-based variation in species sensitivities to environmental gradients. We also built a fifth PGLMM to test a key underlying assumption of phylogenetic community structure: that phylogenetic information serves as a surrogate for trait information about species; this model tests whether the introduction of trait information can explain all variation in species occurrences among communities, leaving no phylogenetic residual variation. We assessed the performance of these PGLMMs using community simulation models and show that PGLMMs have equal or greater statistical power than alternative approaches currently in the literature. Finally, we illustrate the PGLMM advantage of fitting a model to data by showing how variation in species occurrences among communities can be partitioned into phylogenetic and site-specific components, and how fitted models can be used to predict the co-occurrence of phylogenetically related species.

Key words: *ecophylogenetics; environmental gradient; generalized linear models, GLMM; null model; phylogenetic community structure; phylogenetic diversity; phylogenetic signal; trait-based community assembly; trait variation.*

Introduction

The occurrence of species in ecological communities depends on both how species interact with the environment and how they interact with other species. These interactions in turn depend on species-specific traits. For example, the occurrence of a fish species in a lake community might depend on environmental characteristics of the lake (pH, freezing in winter, and so on) and whether the species is tolerant to these characteristics. A species' occurrence might also depend on the presence/absence of other species that either attract the focal species (e.g., prey, mutualists) or repel it (e.g., competitors, predators). For example, a fish species might be excluded from a lake by the presence of a congeneric competitor. Because the occurrence of species depends on traits that dictate their interactions with the environment and other species, and because traits are phylogenetically inherited, we expect to find phylogenetic patterns in the composition of communities, with phylogenetically related species either more or less likely to co-occur at the same sites (Losos 1996, Webb et al. 2002, 2006). We will refer to these general patterns as phylogenetic community structure (Webb et al. 2002).

There is a growing body of literature documenting phylogenetic structure in a wide variety of ecological communities (e.g., Graves and Gotelli 1993, Tofts and Silvertown 2000, Webb et al. 2002, Cavender-Bares et al. 2004, Peres-Neto 2004, Cavender-Bares et al. 2006, Horner-Devine and Bohannan 2006, Weiblen et al. 2006, Helmus et al. 2007*a*, Pillar and Duarte 2010). The typical approach of these studies is to characterize the observed data with some metric of phylogenetic community structure and then perform one or more randomization tests to determine whether the observed value of the metric differs statistically from a null expectation. For example, Webb (2000) examined the tree community composition of 28 sites on Borneo, asking whether co-occurrences of 324 species were random or whether there were phylogenetic patterns. For one of two phylogenetic biodiversity metrics he

Manuscript received 23 June 2010; revised 7 December 2010; accepted 10 December 2010. Corresponding Editor: A. M. Ellison.

[3] E-mail: arives@wisc.edu

considered (the nearest taxa index, NTI), the value for the observed communities differed from the distribution of values generated by permuting species among sites. NTI involves calculating for each species the distance (number of nodes on the phylogenetic tree) to its nearest neighbor, and then computing the average of these minimal distances among species in a community (see also Webb et al. 2002). Among the 28 sites, species co-occurred on average with more closely related species than predicted by chance, where chance is defined by the permutation test. In addition to metrics such as NTI, a similar procedure using randomization tests can be used to compare phylogenetic distance/covariance matrices and species occurrence matrices (e.g., Cavender-Bares et al. 2004, Leibold et al. 2010, Pillar and Duarte 2010).

In contrast to existing approaches that rely on metrics and randomization tests, here we developed statistical models for community structure that can incorporate phylogenetic, environmental, trait, and other information. To clarify the distinction between metrics and models, consider the following simile. Fitting metrics is conceptually similar to performing a randomization test to see if the values of two variables Y and X are associated; this will test for the existence of an association but little more. Fitting a statistical model of community structure is similar to performing regression; fitting the regression model of Y on X gives estimates of coefficients with their confidence intervals, predictions for new values of Y, maximum likelihood values that can be used for model comparisons, diagnostic tests for model goodness of fit, and other statistical information. The approach we developed is essentially a regression for the presence/absence of species among communities that takes the form of a generalized linear mixed model (GLMM; Milner et al. 1999, Krackow and Tkadlec 2001, Kizilkaya and Tempelman 2005, Faes et al. 2006, Gelman and Hill 2007, McCulloch et al. 2008, Bolker et al. 2009).

Phylogenetic GLMMs (PGLMMs) overcome three limitations of the typical metric/randomization approach. First, with the typical approach a single metric summarizes community structure with a single number, and this makes it difficult to explore all but the simplest hypotheses. For the hypothesis (whether phylogenetically related species are more likely to co-occur) tested by Webb (2000), a single metric was adequate (provided it was the right metric: NTI). However, there is growing interest in not just identifying phylogenetic patterns, but also in investigating mechanisms involving environmental factors and species traits (Legendre et al. 1997, Cavender-Bares et al. 2004, McGill et al. 2006, Kraft et al. 2007, Mayfield et al. 2009, Jabot 2010, Pillar and Duarte 2010). For example, we previously attempted to separate the effects of abiotic and biotic factors that generated the co-occurrence patterns of fish species across 890 temperate lakes (Helmus et al. 2007*b*). We first performed regressions to identify environmental factors associated with species occurrences and demonstrated that phylogenetically related species responded similarly to the same environmental factors (i.e., variation in regression coefficients among species for some environmental factors demonstrated phylogenetic signal). After removing the effects of environmental factors, we demonstrated that the residual variation in species occurrences showed phylogenetic repulsion, in which closely related species were less likely to co-occur in the same lake. This investigation could not have been done using simple metrics of phylogenetic community structure. However, each of the steps required separate analyses, leading to the fragmentation of methods and results that did not facilitate the reconstruction of an overall picture of phylogenetic community structure. With a PGLMM, we could perform a single analysis that simultaneously estimates the (potentially phylogenetically determined) sensitivities of species to environmental factors and the (potentially phylogenetically determined) variation in species occurrences that is not explained by these environmental factors.

A second limitation is statistical power. For simple hypotheses, such as that addressed by Webb (2000), analyses based on a simple metric may have reasonable statistical power. However, statistical tests of more complex hypotheses are likely to have low power if the data and analyses are fragmented. For example, Cavender-Bares et al. (2006) investigated the patterns of phylogenetic clustering and phenotypic (trait-based) clustering in the composition of plant communities in Florida. For each pairwise combination of species, they asked whether frequently co-occurring species were more likely to be phylogenetically related, and whether they were more likely to share an assortment of ecological traits. For a taxonomic subset of the species (oaks), they found that those traits shared by species co-occurring in the same communities were less likely to reflect phylogenetic relationships than those traits that were not associated with particular communities; in other words, the phenotypic similarities among species within the same communities were due to convergence rather than phylogenetic relatedness. These conclusions, however, were derived from separate analyses of phylogenies and phenotypes. A PGLMM could provide a single test that uses all available data and should in principle provide greatest statistical power (as a consequence of the Neyman-Pearson lemma; Larsen and Marx 1981:257–261).

A third limitation is that the metric/randomization approach only gives qualitative or yes/no answers about the existence of phylogenetic structure. In contrast, once a PGLMM is fitted, it can be used, for example, to predict the presence of a particular species at a particular site, or predict how frequently two phylogenetically related species likely co-occur. Metrics of phylogenetic community structure are essentially descriptive statistics, whereas PGLMMs are model-based, inferential statistics that attempt to describe the statistical processes underlying observed community patterns (Judge et al. 1985:1). As such, PGLMMs are more flexible, powerful, and informative statistical tools than metrics of community composition.

Below, we illustrate PGLMMs using models designed to address different questions about phylogenetic community structure. These models do not make up an exhaustive list of PGLMMs that can be developed; instead, we chose models that address frequently asked questions in the phylogenetic ecology literature (Webb et al. 2002, Cavender-Bares et al. 2009). We tested these models using simulated data. Simultaneously, we applied alternative metric/randomization tests of phylogenetic community structure. The comparisons between PGLMMs and the alternative approaches reveal the often greater statistical power of PGLMMs. Finally, we illustrate possible information that can be extracted from PGLMMs once they are fitted to data, such as the ability to predict patterns of species occurrences among communities.

Methods

PGLMM

Our formulation of PGLMMs includes environmental variables, species interactions, species traits, and phylogenetic relationships to explain the occurrence (presence/absence) of species in communities. The statistical models we considered have dependent (predicted) variables that take values of 0 (absence) and 1 (presence), fixed effects, and random effects that are used to incorporate specific correlation structures into the model (McCulloch et al. 2008). They can be considered multilevel models that contain both fixed values of coefficients and coefficients that are themselves considered as realizations from a random variable (Gelman and Hill 2007).

We focus on five models that demonstrate the range of problems that PGLMMs can be used to address (Tables 1 and 2). Model I determines whether phylogenetically related species are more likely to occur in the same site. Model II addresses whether phylogenetically related species show similar responses to environmental factors and hence are more likely to co-occur in the same site. Model III includes not only the possibility that species respond to environmental factors in the same way, but also that species show phylogenetic repulsion (sensu Helmus et al. 2007*b*), the pattern in which closely related species are less likely to co-occur in the same site. Model IV tests whether trait values held by species can explain their presence/absence among communities. Finally, model V compares the explanatory powers of trait values of species with their phylogenetic relatedness. Specifically, model V tests the hypothesis that if all ecologically relevant information were known about species traits, then there is no additional information provided by phylogeny. It therefore tests the fundamental assumption that phylogenetic community structure is caused by phylogenetic signal in traits that determine the presence/absence of species in communities; residual phylogenetic community patterns would indicate that some trait(s) or other phylogenetic process (e.g., biogeography) not included in the analysis affects the co-occurrence of species.

Model I: phylogenetic signal in the occurrence of species among sites.—Model I addresses whether phylogenetically closely related species are more likely to co-occur in the same sites. This question has been addressed in numerous studies using numerous methods (e.g., May 1990, Faith 1992, Crozier 1997, Warwick and Clarke 1998, Webb 2000, Cavender-Bares et al. 2004). Although this question is sometimes framed in terms of specific hypotheses about the co-occurrence of species, such as closely related species will likely share similar environmental tolerances and hence occur in the same sites, statistically we only asked whether there is phylogenetic signal in the co-occurrence of species.

To derive an appropriate statistical model, suppose the presence/absence of n species among m sites (communities) is given by the $n \times m$ matrix $\mathbf{\Psi}$ whose jtth element gives the absence or presence of species j at site t. The statistical analysis involves structuring the model so that each element in matrix $\mathbf{\Psi}$ is a dependent datum; specifically, the dependent variable is given by the ($nm \times 1$) vector $\mathbf{Y} = \text{vec}(\mathbf{\Psi})$, where the vec operator stacks consecutive columns of matrix $\mathbf{\Psi}$ on top of each other. The simplest PGLMM we consider is

$$\Pr(Y_i = 1) = \mu_i$$

$$\mu_i = \text{logit}^{-1}(\alpha_{\text{spp}[i]} + b_i + c_{\text{site}[i]})$$

$$\boldsymbol{b} \sim \text{Gaussian}\Big(\mathbf{0}, \text{kron}(\mathbf{I}_m, \sigma^2_{\text{spp}}\mathbf{\Sigma}_{\text{spp}})\Big)$$

$$\boldsymbol{c} \sim \text{Gaussian}(\mathbf{0}, \sigma^2_{\text{site}}\mathbf{I}_m) \qquad (1)$$

where Y_i is the presence (1) or absence (0) of species among sites contained in vector $\mathbf{Y}$. The probabilities μ_i are themselves treated as random variables, with the distribution of logit(μ_i) containing both fixed and random effects. The logit function, logit(μ) = log(μ/[1 − μ]), takes values from $-\infty$ to $+\infty$ as μ varies from 0 to 1.

The model includes as a fixed effect a categorical variable for each species, $\alpha_{\text{spp}[i]}$; here, the function spp[i] gives the identity of the species that corresponds to observation i in the data set (Gelman and Hill 2007:251–2). This fixed effect factors out species-specific differences in prevalence among sites (i.e., different species can occupy a greater or lesser proportion of sites). The random effect b_i accounts for phylogenetic covariances in the co-occurrence of species at a given site and takes a different value for each species–site datum. The $n \times n$ correlation matrix $\mathbf{\Sigma}_{\text{spp}}$ is given by the phylogenetic relationships among species, with the jkth element of $\mathbf{\Sigma}_{\text{spp}}$ determining the phylogenetic correlation in the occurrence of species j and k in the same site. The scalar

TABLE 1. Phylogenetic general linear mixed models (PGLMMs) I–V.

Model	Equation number	Species prevalence	Species occurrence	Species sensitivity	Species repulsion	Species traits	Site richness
I	1	F	P	. . .	. . .	. . .	R
II	2	F	. . .	P, R	. . .	. . .	R
III	3	F	. . .	F	P	. . .	R
IV	7	F	. . .	. . .	. . .	R	R
V	8	F	P	. . .	. . .	R	R

Note: Components of the models are identified as: F, fixed effect; R, random effect; and P, phylogenetic effect. Ellipses indicate that the corresponding components are absent from a model.

σ^2_{spp} is the variance that dictates the overall strength of the phylogenetic covariances in the co-occurrence of species. The Kronecker product kron($\mathbf{I}_m$, $\sigma^2_{\text{spp}}\mathbf{\Sigma}_{\text{spp}}$) generates a ($nm \times nm$) block-diagonal covariance matrix for the $nm \times 1$ random variable $\boldsymbol{b}$ whose diagonal blocks consist of $\sigma^2_{\text{spp}}\mathbf{\Sigma}_{\text{spp}}$, with zeros elsewhere. Finally, the random effect $c_{\text{site}[i]}$ accounts for differences among sites in the numbers of species they contain (i.e., their species richness values). The easiest way to conceptualize $c_{\text{site}[i]}$ is to assume there are m values of c selected from a Gaussian distribution with mean 0 and variance σ^2_{site}, one value for each site. The n values of logit(μ_i) for every species in the same site are increased or decreased by $c_{\text{site}[i]}$, where the function site[i] assigns each datum i to a site.

For Eq. 1 we used the notation of multilevel models (Gelman and Hill 2007: Chapter 12) that displays the structure of the model in terms of each datum i, with fixed and random effects given by Greek and Latin letters, respectively. As a deviation from multilevel notation, however, we present the random effects (or group-level predictors) in vector form to emphasize their multivariate nature. Thus, the m values of $c_{\text{site}[i]}$ are given by the $m \times 1$ random variable $\boldsymbol{c}$, and the nm values of b_i are given by the $nm \times 1$ random variable $\boldsymbol{b}$. Implementation of model I requires the full covariance matrix for $c_{\text{site}[i]}$, which is given by kron($\sigma^2_{\text{site}}\mathbf{I}_m$, $\mathbf{J}_n$), where $\mathbf{J}_\text{n}$ is the $n \times n$ matrix of ones. The full covariance matrix for the model, including both species and site variation, is a block-diagonal matrix having m blocks of dimension $n \times n$ of the following form:

$$\begin{pmatrix} \sigma^2_{\text{sp1,sp1}} + \sigma^2_{\text{site}} & \sigma^2_{\text{sp1,sp2}} + \sigma^2_{\text{site}} & \cdots & \sigma^2_{\text{sp1,sp}n} + \sigma^2_{\text{site}} \\ \sigma^2_{\text{sp1,sp2}} + \sigma^2_{\text{site}} & \sigma^2_{\text{sp2,sp2}} + \sigma^2_{\text{site}} & & \sigma^2_{\text{sp2,sp}n} + \sigma^2_{\text{site}} \\ \vdots & & \ddots & \vdots \\ \sigma^2_{\text{sp1,sp}n} + \sigma^2_{\text{site}} & \sigma^2_{\text{sp2,sp}n} + \sigma^2_{\text{site}} & \cdots & \sigma^2_{\text{sp}n\text{,sp}n} + \sigma^2_{\text{site}} \end{pmatrix}$$

where $\sigma^2_{\text{sp}j\text{,sp}k}$ is the jkth element of $\sigma^2_{\text{spp}}\mathbf{\Sigma}_{\text{spp}}$.

In this model, the overall strength of phylogenetic signal in the co-occurrence of species is determined by σ^2_{spp}, and the overall variation among sites in number of species is determined by σ^2_{site}. Because the variance of a binary process with expectation μ is $\mu(1 - \mu)$, the variances in $\mathbf{Y}$ are determined solely by the expectations μ_i, and there is no "unexplained" variance in logit(μ_i) (Gelman and Hill 2007: Chapter 5). The addition of a random variable for unexplained variation in logit(μ_i) would have indistinguishable effects from decrementing

TABLE 2. Summary of simulation models used to test the performance of statistical models I–V.

Model	Simulation model	Statistical model	Test
I	two environmental variables; phylogenetic patterns in sensitivity to two gradients	zero environmental variables; phylogenetic patterns in co-occurrence	detect phylogenetic patterns in co-occurrence
II	two environmental variables; phylogenetic patterns in sensitivity to two gradients	one environmental variable; phylogenetic patterns in co-occurrence; species-specific and phylogenetic patterns in response to a single environmental gradient	detect phylogenetic patterns in species sensitivity to an environmental gradient
III	one environmental variable ; phylogenetic repulsion	one environmental variable; phylogenetic repulsion	separate phylogenetic repulsion from sensitivity to an environmental gradient
IV	one environmental variable; phylogenetic patterns in sensitivity to a gradient	zero environmental variables; trait variation among species	detect trait-based patterns in co-occurrence
V	one or two environmental variables; phylogenetic patterns in sensitivity to gradients	zero environmental variables; trait variation among species; phylogenetic patterns in co-occurrence	distinguish trait-based from phylogenetic effects on co-occurrence
All	multiplicative survival, e.g., see Eq. 9 and following: $P_{jt} = p_{x,jt}p_{y,jt}p_{s,t}$	generalized linear mixed model (GLMM), e.g., see Eq. 1: $\mu_i = \text{logit}^{-1}(\alpha_{\text{spp}[i]} + b_i + c_{\text{site}[i]})$	

the variances σ^2_{spp} and σ^2_{site}, thereby introducing an identifiability problem in which it is impossible to estimate the unexplained variation independently of σ^2_{spp} and σ^2_{site}. In general, care needs to be taken in constructing models so that all the parameters capture distinct properties of the data; when this is not the case, parameter estimates can be highly variable and strongly correlated with each other, the symptoms of non-identifiability.

The choice of which factors to include as fixed vs. random effects needs to be made strategically. The factor b_i is necessarily a random effect so that it is possible to incorporate the phylogenetic covariance matrix. We included $c_{\text{site}[i]}$ as a random effect because the nonzero elements of the covariance matrices associated with b_i and $c_{\text{site}[i]}$, kron($\mathbf{I}_m$, $\sigma^2_{\text{spp}}\mathbf{\Sigma}_{\text{spp}}$) and kron($\sigma^2_{\text{site}}\mathbf{I}_m$, $\mathbf{J}_n$), respectively, perfectly coincide (see matrix two paragraphs above). This occurs because b_i contains information about the co-occurrence of phylogenetically related species in the same site, and $c_{\text{site}[i]}$ contains information about whether more species overall (regardless of phylogenetic relationships) occur in the same site. The reason for including $c_{\text{site}[i]}$ as a random effect is that otherwise variation in the number of species per site could cause the false identification of phylogenetic patterns; that is, species may be more likely to co-occur in sites due to site-to-site variation rather than phylogenetic relationships. We incorporated the variation in prevalence among species, $\alpha_{\text{spp}[i]}$, as a fixed effect; it could be treated as a random effect, although in application we have found that minimizing the number of random effects in the model generally leads to better estimates of those random effects that are included. By including $\alpha_{\text{spp}[i]}$ in the model, phylogenetic patterns in the co-occurrence of species captured by $\sigma^2_{\text{spp}}\mathbf{\Sigma}_{\text{spp}}$ will not reflect phylogenetic patterns observed in species-specific mean incidences among sites.

To derive a form of the phylogenetic covariance matrix $\sigma^2_{\text{spp}}\mathbf{\Sigma}_{\text{spp}}$, we assume that the occurrence of species is governed by a hypothetical (unmeasured) trait that evolves up the phylogenetic tree. If evolution follows a Brownian motion process, in which incremental increases and decreases in the trait value occur at random, the distribution of trait values of species at the tips of the phylogenetic tree is multivariate Gaussian; the covariance matrix $\sigma^2_{\text{spp}}\mathbf{\Sigma}_{\text{spp}}$ has *jk*th elements that are proportional to the shared branch length between species *j* and *k*, with the diagonal elements equaling one. Variation in this hypothetical trait generates variation in μ_i that then affects the occurrence of species among sites. Using other evolutionary models to derive $\sigma^2_{\text{spp}}\mathbf{\Sigma}_{\text{spp}}$ is possible, such as the Ornstein-Uhlenbeck (OU) stabilizing selection model (e.g., Blomberg et al. 2003); however, there is an identifiability problem with using the OU model since it depends on a parameter governing the strength of stabilizing selection that cannot be distinguished from σ^2_{spp} when fitting presence/absence data. In our formulation of $\sigma^2_{\text{spp}}\mathbf{\Sigma}_{\text{spp}}$, model I can only detect communities with a phylogenetically underdispersed structure (i.e., closely related species generally co-occur). With model III we describe an approach to detect the opposite structure, phylogenetic overdispersion.

Model II: phylogenetic signal in the sensitivity of species to an environmental factor.—Phylogenetically closely related species are often assumed to be ecologically similar, and therefore, if there is a set of strong environmental drivers determining the distribution of species, phylogenetically related species that share common responses to the environmental drivers should be more likely to co-occur. To address this hypothesis, one approach is to calculate phylogenetic differences in the species composition of communities and associate these differences with differences in environmental factors. For example, Hardy and Senterre (2007) correlated a phylogenetic measure of differences between tree communities (beta diversity) in Equatorial Guinea with environmental differences between sites and found that greater compositional differences between communities were associated with greater differences in altitude. In contrast to this approach, the model we developed here uses data on environmental differences between sites to investigate directly whether phylogenetically related species respond to environmental factors in similar ways. The analysis can be envisioned as a set of logistic regressions of species presence/absence on an environmental variable, where these regressions are performed simultaneously for all species in a species pool. The model then tests whether there is phylogenetic signal in the variation in logistic regression coefficients among species (Helmus et al. 2007*b*).

This model has the following form:

$$\Pr(Y_i = 1) = \mu_i$$

$$\mu_i = \text{logit}^{-1}(\alpha_{\text{spp}[i]} + b_{\text{spp}[i]}x_{\text{site}[i]} + c_{\text{site}[i]})\sigma^2_\alpha$$

$$b_{\text{spp}[i]} = \beta + e_{\text{slope}[i]} + g_{\text{phylo}[i]}$$

$$\boldsymbol{e} \sim \text{Gaussian}(0, \sigma^2_{\text{slope}}\mathbf{I}_n)$$

$$\boldsymbol{g} \sim \text{Gaussian}(0, \sigma^2_{\text{phylo}}\mathbf{\Sigma}_{\text{spp}})$$

$$\boldsymbol{c} \sim \text{Gaussian}(0, \sigma^2_{\text{site}}\mathbf{I}_m). \qquad (2)$$

The difference from model I is that logit(μ_i) is assumed to depend linearly on the environmental factor x through the regression coefficient $b_{\text{spp}[i]}$. The regression coefficient itself is assumed to be a Gaussian random variable with mean β and covariance matrix given the sum of $\sigma^2_{\text{slope}}\mathbf{I}_n$ and $\sigma^2_{\text{phylo}}\mathbf{\Sigma}_{\text{spp}}$. The first of these covariance matrices accounts for species-specific differences that are not phylogenetically related among species, while the second accounts for phylogenetic relatedness assuming that covariances among species are determined by a Brownian motion model of evolution. For ease in interpretation, we have assumed the environmental variable x is centered on zero, so that differences in $\alpha_{\mathrm{spp[i]}}$ among species correspond to species-specific mean incidence among sites.

To implement model II, we need the covariance matrix for $b_{\mathrm{spp}[i]}x_{\mathrm{site}[i]}$. This is given by $\mathbf{\Sigma}_{\mathrm{slope}} = \mathrm{diag}(\mathbf{X}) \times \mathrm{kron}(\mathbf{J}_m, \sigma^2_{\mathrm{slope}}\mathbf{I}_n + \sigma^2_{\mathrm{phylo}}\mathbf{\Sigma}_{\mathrm{spp}}) \times \mathrm{diag}(\mathbf{X})$, where $\mathbf{X}$ is the $(nm \times 1)$ vector containing values of x_i, diag() represents the $(nm \times nm)$ diagonal matrix whose diagonal elements are $\mathbf{X}$, and × denotes matrix multiplication. The $nm \times nm$ covariance matrix $\mathbf{\Sigma}_{\mathrm{slope}}$ incorporates information about both the phylogenetic correlations among species, $\mathbf{\Sigma}_{\mathrm{spp}}$, and the environmental factor, $\mathbf{X}$.

In this model, we assume that phylogenetic patterns of species occurrences are caused by related species responding similarly to the environmental factor x. Unlike Eq. 1, there is no random effect corresponding to species alone. This is a strategic decision driven by the statistical difficulty of separating phylogenetic effects that operate through sensitivities to an environmental factor from those that do not. When confronted with the problem of distinguishing these two sources of phylogenetic pattern in real data sets, such as when species interactions cause correlations among closely related species that are independent of any environmental factor, a reasonable approach is to fit models with $b_{\mathrm{spp}[i]}x_{\mathrm{site}[i]}$ (as in model II) or with b_i (as in model I), and then compare the fit of the two models.

Model III: phylogenetic attraction and repulsion.—The third model we consider assumes that phylogenetically closely related species might show similar responses to an environmental factor (leading to "phylogenetic attraction," Helmus et al. 2007*b*), yet after factoring out this attraction phylogenetically closely related species tend not to co-occur (leading to "phylogenetic repulsion"; Helmus et al. 2007*b*). This scenario could arise if there were simultaneous environmental filtering, with related species responding similarly to the filter, and competition in which related species are more likely to exclude each other.

The model has the following form:

$$\begin{aligned}
&\Pr(Y_i = 1) = \mu_i\\
&\qquad \mu_i = \mathrm{logit}^{-1}(\alpha_{\mathrm{spp}[i]} + \beta_{\mathrm{spp}[i]}x_{\mathrm{site}[i]} + c_{\mathrm{site}[i]} + d_i)\\
&\qquad \boldsymbol{c} \sim \mathrm{Gaussian}(0, \sigma^2_{\mathrm{site}}\mathbf{I}_m)\\
&\qquad \boldsymbol{d} \sim \mathrm{Gaussian}\Big(0, \mathrm{kron}(\mathbf{I}_m, \sigma^2_{\mathrm{repulse}}\mathbf{\Sigma}_{\mathrm{repulse}})\Big). \qquad (3)
\end{aligned}$$

Many terms are identical to the preceding models. In contrast to model II, however, species-specific differences in the slopes in response to the environmental factor x are treated as a fixed effect. The random effect d_i gives the effect of phylogenetic repulsion on species co-occurrence contained within the correlation matrix $\mathbf{\Sigma}_{\mathrm{repulse}}$.

To implement this model that includes phylogenetic repulsion, we need a biologically sensible covariance matrix $\mathbf{\Sigma}_{\mathrm{repulse}}$. For simplicity, we used $\sigma^2_{\mathrm{repulse}}\mathbf{\Sigma}_{\mathrm{repulse}} = (\sigma^2_{\mathrm{spp}}\mathbf{\Sigma}_{\mathrm{spp}})^{-1}$; this formulation guarantees that $\sigma^2_{\mathrm{repulse}}$ $\mathbf{\Sigma}_{\mathrm{repulse}}$ is a legitimate covariance matrix (i.e., is positive definite). In this formulation, the strength of repulsion between any two species is proportional to their phylogenetic relatedness. Theoretical justification for this formulation is given in Appendix A.

Model IV: trait-based community structure.—In models I–III phylogenetic information was used implicitly as a surrogate for information about species traits that dictate their sensitivities to biotic or abiotic forces. Here, we considered the case in which there is information for each species about traits that determine the sensitivities of species to an unknown environmental factor. We were not interested in whether there are phylogenetic patterns in the occurrence of species among sites, because we assume that any phylogenetic signal is absorbed by information about species traits; phylogenetic signal could occur in the distribution of traits among species, but because we assume that these trait values are known, this phylogenetic signal gives no additional information. Although we know trait values, we assume that we do not know what environmental factors drive the trait-based pattern of presence/absence of species at sites.

To build a statistical model, we needed to derive an expected covariance matrix for the co-occurrence of species based upon their known trait values but with no knowledge of the value of environmental factors. Here, we selected a simple though realistic derivation. We assume that for a given species j, the probability that it occurs in a site depends on an unmeasured environmental factor x:

$$\mathrm{logit}(\mu_i) = \alpha_i + \beta_{\mathrm{spp}[i]}x_i + e_i \qquad (4)$$

where $\beta_{\mathrm{spp}[i]}$ is the known (i.e., fixed) trait-based sensitivity of species to the environmental factor, and e_i is a Gaussian random variable with expectation zero. If we assume values of $\beta_{\mathrm{spp}[i]}$ are standardized to have mean zero, then the covariance between the occurrence of two species is

$$\mathrm{cov}\Big(\mathrm{logit}(\mu_i), \mathrm{logit}(\mu_{i'})\Big) = \beta_{\mathrm{spp}[i]}\beta_{\mathrm{spp}[i']}\ \mathrm{var}(x) \qquad (5)$$

where x is an unknown value of the environmental random variable. Thus, a reasonable covariance matrix for species based on their trait values is

$$\sigma^2_{\mathrm{trait}}\mathbf{\Sigma}_{\mathrm{trait}} = \sigma^2_{\mathrm{trait}}(\boldsymbol{\beta\beta}') \qquad (6)$$

where $\boldsymbol{\beta}$ is the $n \times 1$ vector of values of β. Technically, $\mathbf{\Sigma}_{\mathrm{trait}}$ is not a well-defined covariance matrix because it is not positive definite; nonetheless, it is positive semidefinite, and therefore, can be used without logical or computational difficulties.

The overall statistical model can be formulated similarly to model I as follows:

$$\Pr(Y_i = 1) = \mu_i$$

$$\mu_i = \text{logit}^{-1}(\alpha_{\text{spp}[i]} + c_{\text{site}[i]} + f_i)$$

$$\boldsymbol{c} \sim \text{Gaussian}(0, \sigma^2_{\text{site}}\mathbf{I}_m)$$

$$\boldsymbol{f} \sim \text{Gaussian}\Big(0, \text{kron}(\mathbf{I}_m, \sigma^2_{\text{trait}}\boldsymbol{\Sigma}_{\text{trait}})\Big). \tag{7}$$

This model assumes that the covariance in the occurrence of species through $\sigma^2{}_{\text{trait}}\boldsymbol{\Sigma}_{\text{trait}}$ is proportional to the product of their standardized trait values.

Model V: trait-based vs. phylogenetic community structure.—In model IV we assume that all information about the similarity among species important for their co-occurrence in communities is captured by measured trait values. Alternatively, we could assume that beyond trait values, there is also information provided by their phylogenetic relationships (e.g., Cavender-Bares et al. 2006). This might occur, for example, if in addition to measured traits among species, there are unmeasured traits that show phylogenetic patterns or there is some type of phylogeographic structure in the community data set. In this case, there may be simultaneous effects of trait values and phylogeny on community composition.

The appropriate model that includes both trait-based and phylogenetic processes determining the presence/absence of species among communities is a merger of models I and IV:

$$\Pr(Y_i = 1) = \mu_i$$

$$\mu_i = \text{logit}^{-1}(\alpha_{\text{spp}[i]} + b_i + c_{\text{site}[i]} + f_i)$$

$$\boldsymbol{b} \sim \text{Gaussian}\Big(0, \text{kron}(\mathbf{I}_m, \sigma^2_{\text{spp}}\boldsymbol{\Sigma}_{\text{spp}})\Big)$$

$$\boldsymbol{c} \sim \text{Gaussian}(0, \sigma^2_{\text{site}}\mathbf{I}_m)$$

$$\boldsymbol{f} \sim \text{Gaussian}\Big(0, \text{kron}(\mathbf{I}_m, \sigma^2_{\text{trait}}\boldsymbol{\Sigma}_{\text{trait}})\Big). \tag{8}$$

This model serves as a test of the hypothesis that, if full information were known about species traits, then phylogenetic information would not provide additional information about species occurrences in communities. In other words, phylogenetic information serves only as a surrogate for trait information.

Estimation.—There are various approaches that, in principle, can be used for parameter estimation in PGLMMs, although none is easy to implement (Bolker et al. 2009). For example, covariance matrices cannot be specified in the GLMM R package lmer (R Development Core Team 2005, Bates et al. 2008), although this limitation might be overcome in future versions of lmer (D. Bates, *personal communication*). Both WinBUGS and the MCMCglmm R package (Hadfield 2010) estimate parameters with Markov Chain Monte Carlo (MCMC) approaches (Gelman and Hill 2007). However, specialized programming is required for both, and we found our models difficult to implement.

Our estimation approach combines penalized quasi-likelihood (PQL) and restricted maximum likelihood (REML) in a two-step process. Details are presented in Appendix B, along with extensive numerical explorations investigating the statistical properties of the approach. In brief, estimation is performed iteratively. First, the coefficients for fixed effects are estimated via PQL conditional on the working estimates of the variances of the random effects. Second, the variances of the random effects are recalculated using REML with the estimates conditional on the updated estimates for the fixed effects. These steps are sequentially iterated to convergence. For statistical inference about the variances in the random effects σ^2, we used the profile restricted likelihood conditional on the PQL estimates of the coefficients of the fixed effects. For moderately sized problems (number of species–sites $nm < 2000$) our approach is fast and robust. MATLAB (MathWorks 2005) code is provided for this procedure in the online Supplement, which has been translated into the R statistical language in the package Picante (Kembel et al. 2010).

Performance tests

We assessed the performance of our PGLMMs using simulated data sets that contain patterns that phylogenetic community methods are designed to identify. To compare with the performance of the PGLMMs, we used alternative "standard" approaches from the literature that aim to identify the same patterns. We tuned the simulations so that the patterns were weak, thereby testing the statistical power of PGLMMs against alternative approaches. Table 2 summarizes the simulation models and tests.

Phylogenetic community assembly simulations.—The community assembly simulations we used to investigate the performance of PGLMMs contain environmental gradients, trait-based sensitivities to these gradients, and phylogenetic repulsion, as appropriate for the different tests. The simulation models differ from the statistical models I–V in both structure (how presence/absence of species are modeled) and often in terms of the number of environmental variables included (Table 2). For all simulations, we assume that there are 31 sites (communities) and 32 species. While phylogenetic tree shape may affect the power of PGLMMs to detect phylogenetic patterns as it does for the standard phylogenetic ecology approaches (Kraft et al. 2007, Swenson 2009), assessing this effect is beyond the scope of the work we present here. In practice, the power of a PGLMM should be assessed for the particular tree that is used with a

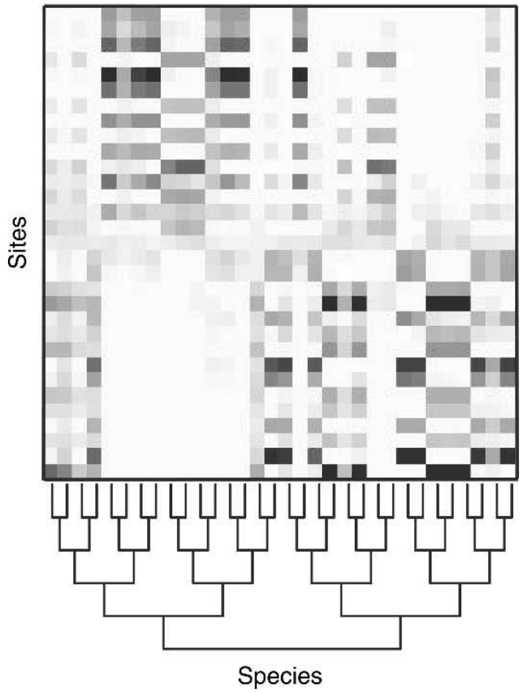

FIG. 1. Probability of occurrence of 32 species among 31 sites from a simulation model (darker indicates greater probability). Species are phylogenetically related according to a fully balanced tree. Sites are sorted by one of two simulated environmental gradients, and phylogenetic patterns in species sensitivities to the environmental gradients are seen in the similar distributions of related species among sites.

particular data set. For simplicity we used a fully balanced tree (Fig. 1).

The response of species to an environmental gradient is given by the inverse logit function:

$$p_{x,jt} = q_t \frac{\exp(a + b_j x_t)}{1 + \exp(a + b_j x_t)} \tag{9}$$

where $p_{x,jt}$ is the probability of species j occurring at site t (as determined by the environment gradient x), x_t is the value of the environmental factor at site t (centered so that the mean value of x_t is zero), and a and b_j are coefficients. We assume that a takes the same value for each species, whereas the sensitivity of species to the environmental gradient, b_j, takes species-specific values. In particular, b_j is assumed to be a trait that evolves in a Brownian motion fashion up the phylogenetic tree and therefore has a Gaussian distribution with covariance matrix $\sigma^2_{\mathrm{spp}}\boldsymbol{\Sigma}_{\mathrm{spp}}$. We also impose a reduction in the probability of species occurring in a site using the random variable q_t that is selected for each site t from a uniform distribution between 0.5 and 1. This simulates variation among sites in species richness values that is independent of the environmental gradients. The overall probability that species j occurs in site t having values x_t and y_t for two environmental factors x and y is $P_{jt} = p_{x,jt}p_{y,jt}$. Thus, components that give the probabilities of occurring in a site as determined by different environmental gradients or site-specific factors are combined multiplicatively to give the overall probability that a species occurs in a given site. This contrasts PGLMMs I–V in which all independent variables are contained within a logit^{-1} function.

An example of simulated values of P_{jt} is given in Fig. 1. When analyzing the data only for the existence of phylogenetic patterns in the occurrence of species (models I, IV, and V), we removed any sites in the simulated data sets with no species. In contrast, for models II and III that include environmental information, we did not remove any sites, because even sites with no species still provide information about the effect of the environmental factor on species occurrences. In all models we only considered simulations in which all species occurred in at least one site.

To incorporate repulsion among phylogenetically related species in simulations for model III, we assume that the probability of species j occurring in site t is given by

$$p_{r,jt} = \frac{\exp(c + d_j)}{1 + \exp(c + d_j)} \tag{10}$$

where the species-specific values d_j are assumed to have a Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}_{\mathrm{repulse}}$ in a manner similar to model III (Eq. 3). The overall probability that species j occurs at site t having value x_t for the environmental factor x is $P_{jt} = p_{x,jt}p_{r,jt}$.

For simulations to assess models III and IV, there is a single environmental factor affecting community composition, whereas for the other models there are two. For model III we wanted to statistically extract all environmental variation that caused phylogenetic attraction in order to expose phylogenetic repulsion. For model IV we wanted to determine whether community composition could be fully explained by trait-based responses to the environment. We therefore only included one environmental factor in the simulations for these models. For the other models, we used two environmental factors to allow a residual phylogenetic pattern even after removing the effects of a single environmental factor. For all models, the values of environmental factors were evenly spaced across sites, and when there were two factors they were assigned independently of each other. Thus, the sites can be considered to fall along one (models III, IV) or two unrelated (models I, II, and V) gradients.

Scaling parameters in the simulation models were selected so that the average numbers of sites occupied per species were 4.3 ± 1.7 (mean ± SD; models I, II, and V), 7.8 ± 2.1 (model III), and 11.6 ± 3.7 (model IV). There were averages of 4.8 ± 2.0 (models I, II, and V), 8.0 ± 2.0 (model III), and 12.0 ± 3.5 species per site (model IV). Thus, the simulations generate relatively small communities that should challenge the statistical techniques to find phylogenetic patterns. The MATLAB code in the online Supplement gives details of the simulations and parameter values.

Alternative to model I: phylogenetic signal in the occurrence of species among sites.—As an alternative to model I, we tested for the presence of phylogenetic signal in the occurrence of species among sites using the following procedure. For each of the 31 sites in the simulated data sets, we computed the mean pairwise phylogenetic nodal distance between species (mean phylogenetic distance, MPD; Webb 2000), and then took the average MPD among all sites. To obtain a statistical test for phylogenetic signal, we generated 10 000 permutation data sets by permuting species among sites, thereby preserving the expected number of sites occupied by each species. We then calculated the average MPD value for each randomized data set. Because under the null hypothesis the value of MPD is independent of the number of species in sites, we did not use a permutation algorithm designed to maintain the same number of species per site. Significant phylogenetic signal was determined if the average MPD of the original data set fell below the 2.5% quantile of the permutation distribution of the 10 000 null average MPD values.

Alternative to model II: phylogenetic signal in the sensitivity of species to environmental variation.—To compare with model II, we used the approach of Helmus et al. (2007*b*). For each simulation data set, we fit a logistic regression (with a Firth correction) of the presence/absence of each species in the data set on one of the environmental factors. We then tested for phylogenetic signal in the coefficients of these regressions with a phylogenetic linear regression using the MATLAB code RegressionV2.m (Lavin et al. 2008). RegressionV2.m fits data assuming that the residual variation follows an Ornstein-Uhlenbeck process of evolution. This gives an estimate of the parameter $d \geq 0$ that takes a value of 0 if there is no phylogenetic signal and 1 if the signal is that predicted by a Brownian motion model of evolution. Statistical significance of the null hypothesis of no signal was tested using bootstrapping. Programs for these calculations are available in the R package Picante (Kembel et al. 2010).

Alternative to model III: phylogenetic attraction and repulsion.—As an alternative to model III, we again used the approach of Helmus et al. (2007*b*). First, we factored out the effect of the environmental gradient by regressing the presence/absence of each species on the environmental factor as done for model II. We then computed standardized residuals from each logistic regression:

$$\rho_{jt} = \frac{y_{jt} - \mu_{jt}}{[\mu_{jt}(1-\mu_{jt})]^{1/2}} \tag{11}$$

where

$$\mu_{jt} = \frac{\exp(b_{0,j} + b_{1,j}x_t)}{1 + \exp(b_{0,j} + b_{1,j}x_t)}$$

and y_{jt} is the presence (1) or absence (0) of species j at site t, $b_{0,j}$ and $b_{1,j}$ are the species-specific logistic regression coefficients, and x_t is the value of the environmental factor at site t. For each pair of species, the covariance between r_{jt} values is a measure of how likely it is that the species co-occur after the environmental effect is removed. We correlated these covariances against species pairwise phylogenetic covariances given by $\mathbf{\Sigma}_{\text{spp}}$. For each simulation data set, this observed correlation value was compared to the distribution of values constructed by permuting r_{jt} among sites and recomputing the correlations 10 000 times. Significant phylogenetic repulsion is determined if the resulting correlation c fell below the 2.5% quantile of the permutation distribution.

Alternative to model IV: trait-based community structure.—Model IV investigates whether community composition can be explained by the traits exhibited by species. An alternative method is to use a metric of community trait spacing (Cornwell and Ackerly 2009 and references therein). For simulated data sets we computed the standard deviation in trait values shared by species in the same site and then averaged the values across all sites. For a statistical test, we permuted the species among sites 10 000 times to construct a distribution of average standard deviations in trait values under the null hypothesis that species (and hence trait) distributions were independent of site (Cornwell and Ackerly 2009). Statistically significant trait-based community structure was identified if the observed average standard deviation in trait values fell below the 2.5% quantile of the permutation distribution.

Alternative to model V: trait-based vs. phylogenetic community structure.—Model V addresses whether information about species trait values is sufficient to explain phylogenetic community structure or whether even after incorporating trait information there is residual phylogenetic structure. We know of no alternative method to PGLMM that can be used to address this question. Therefore, we compared the results of model V to those from model I applied to the same simulated data sets. Model I identifies phylogenetic structure without incorporating trait information, whereas model V is identical to model I except trait information is incorporated. Comparing the two demonstrates how much information about community composition can be extracted from trait information.

Results

For each PGLMM and corresponding alternative, we simulated 100 data sets under conditions that give relatively weak phylogenetic patterns. The PGLMMs all had equal or greater power to detect phylogenetic patterns than alternative tests (Table 3). To investigate whether this was due to PGLMMs giving false positives (Type I errors), we repeated the simulations under the assumption that species were phylogenetically unrelated. Both PGLMMs and the alternatives give close to the

TABLE 3. Number of simulation runs out of 100 showing statistically significant phylogenetic patterns in community composition ($\alpha = 0.025$, one-tailed).

		Simulated with phylogenetic signal		Simulated without phylogenetic signal	
Model	Description	PGLMM†	Alternative	PGLMM†	Alternative
I	species co-occurrence	87	73	5	3
II	response to environmental gradient	83	27	7	0
III	phylogenetic repulsion	53	0	1	0
IV	trait-based community structure	64	19	1	4

† Phylogenetic generalized linear mixed model.

expected false positive rates at the nominal $\alpha < 0.025$ (one-tailed) level (bottom half of Table 3). Bootstrap simulations to analyze the statistical properties of the PGLMM estimator are given in Appendix B.

We will discuss the performance of each model separately. We then highlight other analyses that are possible once a PGLMM is fit to data (analyses that are not possible using metric/randomization methods). In particular, we illustrate how PGLMMs partition sources of variance in community composition and how PGLMMs predict the occurrence of species from the presence/absence of phylogenetically related species.

Identification of community structure

Model I: phylogenetic signal in the occurrence of species among sites.—In 87% of the data sets, model I identified values of σ_{spp} statistically significantly greater than zero (Fig. 2). The alternative test using mean phylogenetic distances among species (MPD) performed almost as well, rejecting the null model of random community assembly in 73% of the data sets (Table 3). As expected, there is a negative correlation between σ_{spp} and standardized average MPD values across communities (Fig. 2), because larger σ_{spp} and smaller average MPD correspond to greater phylogenetic community structure.

Model II: phylogenetic signal in the sensitivity of species to an environmental gradient.—For simulations in which the distribution of species was determined by two environmental factors, and one of these factors was used to predict the patterns of species occurrences among sites, 83% of the simulated data sets were identified by model II (Eq. 2) as having statistically significant phylogenetic signal, $\sigma_{phylo} > 0$ (Fig. 3). In contrast, the estimates of phylogenetic signal, d, from the

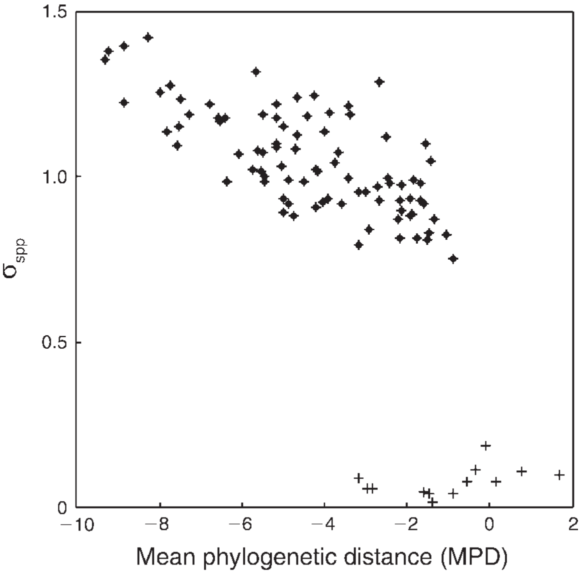

FIG. 2. PGLMM I (phylogenetic generalized linear mixed model) vs. an alternative test: for 100 simulation data sets, the average mean phylogenetic distance (MPD) values and the estimates of the measure of phylogenic signal, σ_{spp}, from model I (Eq. 1). The average MPD values are standardized using the permutation distribution so that under the null hypothesis of no phylogenetic signal the expectation is 0 and the variance is 1. Four-pointed stars give values of σ_{spp} that are statistically different from 0 at the $\alpha < 0.025$ level as obtained by profile likelihoods, and +'s denote nonsignificant values.

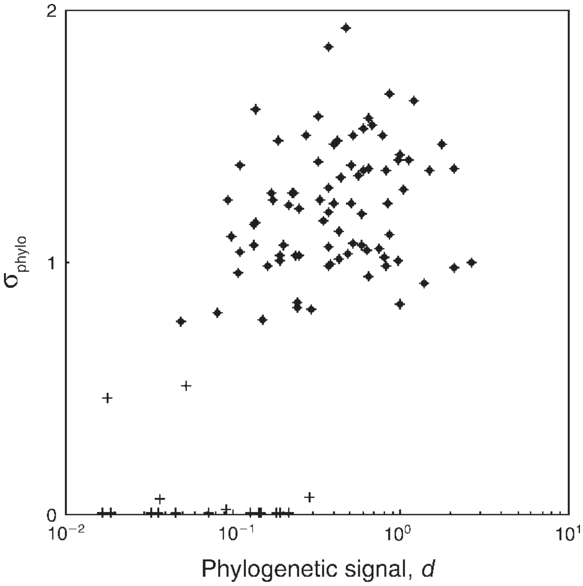

FIG. 3. PGLMM II vs. an alternative test: for 100 simulation data sets, phylogenetic variation in species-specific sensitivities to one of two environmental factors, σ_{phylo} (model II, Eq. 2), vs. phylogenetic signal estimates, d, in the values of logistic regression slopes. Four-pointed stars give values of σ_{phylo} that are statistically different from 0 at the $\alpha < 0.025$ level as obtained by profile likelihoods, and +'s denote nonsignificant values.

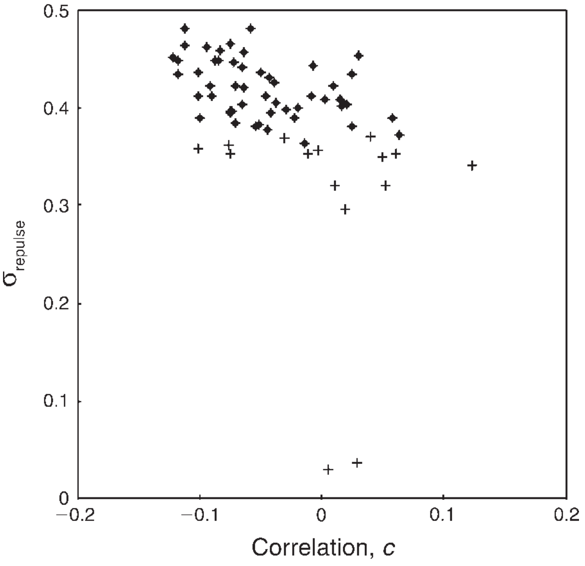

FIG. 4. PGLMM III vs. an alternative test: for 100 simulation data sets, phylogenetic repulsion in the co-occurrence of species, σ_{repulse}, when accounting for the effect of an environmental gradient (model III, Eq. 3) vs. the correlation between the covariances in species standardized residuals and phylogenetic covariances, *c*. Values of *c* are standardized using the permutation distribution so that under the null hypothesis of no phylogenetic signal the expectation is 0 and the variance is 1. Four-pointed stars give values of σ_{repulse} that are statistically different from 0 at the $\alpha < 0.025$ level as obtained by profile likelihoods, and +'s denote nonsignificant values.

alternative method were only statistically significant for 27% of the simulated data sets. Also, there was only a weak positive relationship between σ_{phylo} and *d*.

Model III: phylogenetic attraction and repulsion.—Model III accounts for the effect of environmental gradients that generate positive co-occurrences between phylogenetically related species, and identifies phylogenetic repulsion as residual negative co-occurrence patterns that could result if phylogenetically closely related species were more likely to exclude each other. Model III identified statistically significant phylogenetic repulsion ($\sigma_{\text{repulse}} > 0$) in 53 of 100 simulated data sets (Table 3). In contrast, the alternative method identified statistically significant phylogenetic signal in none of the data sets (Table 3). There was a negative correlation between estimates of σ_{repulse} calculated from model III and the correlations *c* between residual covariances and phylogenetic covariances (Fig. 4), although the statistical power of the alternative method based on *c* was much lower. Helmus et al. (2007*b*) found statistically significant repulsion in a fish community data set, although it contained data from 890 lakes, which presumably made up for the apparently low statistical power of this test.

Model IV: trait-based community structure.—Model IV is designed to identify whether information about trait values of species can explain species co-occurrences among communities. In 64 of 100 simulation data sets (Fig. 5) model IV identified trait-based patterns in community composition ($\sigma_{\text{trait}} > 0$, $\alpha = 0.025$, one-tailed). In comparison, 19 of the data sets were identified as having trait-based signal by the alternative method that depends on the variation in trait values among species within the same community (e.g., Cornwell and Ackerly 2009).

Model V: trait-based vs. phylogenetic community structure.—The use of phylogenetic information in assaying community structure is often justified as a surrogate for trait-based information when species traits are unknown (e.g., Webb 2000). To investigate this proposition, we compared model V that contains both trait and phylogenetic information with model I that contains only phylogenetic information. Neither model uses information about the environmental factor at each site; if this were available, then it could be used to relate species traits to environmental conditions as in model II. Model V identified trait-based patterns in all simulated data sets, while it identified residual phylogenetic patterns in only 1 (Table 4). Model I identified phylogenetic patterns in 85 of the simulation data sets. Therefore, inclusion of trait information in model V accounts for almost all of the phylogenetic patterns that were identified by model I. In other words, if there is information about species traits that determines the composition of communities, then all phylogenetic information is captured by trait information.

In a second simulation study, we considered two environmental gradients. For model V we assumed only trait values associated with one of the environmental factors are known. In this case, model V identified trait-based patterns in all of the 100 simulated data sets, and identified phylogenetic patterns in 47 (Table 4). Finding

TABLE 4. Number of simulation data sets out of 100 showing statistically significant trait and phylogenetic patterns in community composition as assessed by models I and V ($\alpha = 0.025$, one-tailed).

	One environmental factor		Two environmental factors	
Pattern	Model I	Model V	Model I	Model V
Traits ($\sigma_{\text{trait}} > 0$)	···	100	···	100
Phylogeny ($\sigma_{\text{spp}} > 0$)	85	1	87	47

Notes: For model I, no trait value was incorporated (indicated by ellipses). For model V species sensitivities were known for only one environmental factor.

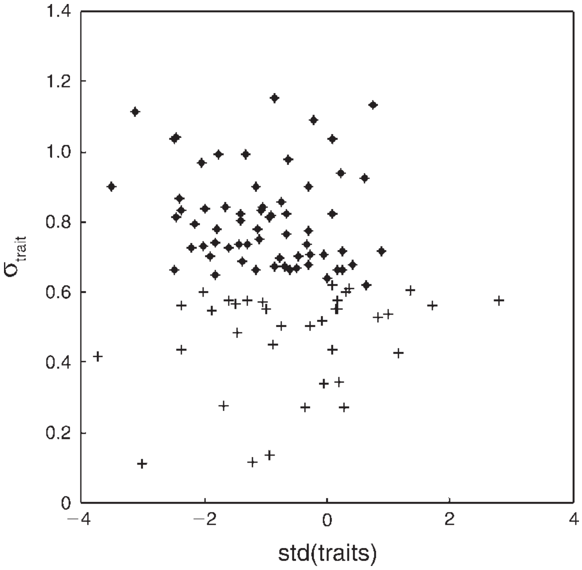

FIG. 5. PGLMM IV vs. an alternative test: for 100 simulation data sets, the trait-based pattern in the co-occurrence of species, σ_{trait} (model IV, Eq. 7), vs. the average standard deviation in trait values among species in the same communities, std(traits). Values of std(traits) are standardized using the permutation distribution so that under the null hypothesis of no phylogenetic signal the expectation is 0 and the variance is 1. Four-pointed stars give values of σ_{trait} that are statistically different from 0 at the $\alpha < 0.025$ level as obtained by profile likelihoods, and +'s denote nonsignificant values.

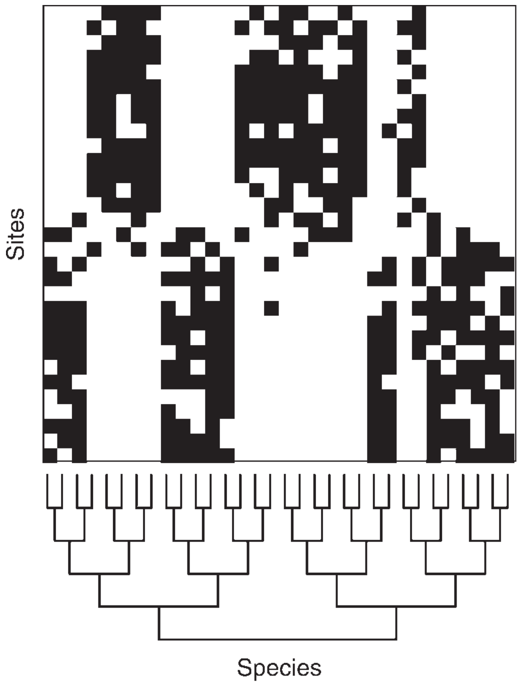

FIG. 6. Simulation data set used to estimate model 1 (Table 5). The occurrence (black squares) of species 1 is given in the leftmost column, and the occurrences at site 1 are given by the lowest row (see *Uses of fitted models* in *Results*).

TABLE 5. A partial summary of statistical results from PGLMM I (see Eq. 1) fit to a simulated data set.

			Explained variance (%)	
Parameter	Estimate	95% CL	Random effects	All effects
σ^2_{spp}	0.92	(0.48, 1.47)	81	71
σ^2_{site}	0.21	(0, 0.65)	19	16
(σ^2_{α})†	0.16			13

Note: The restricted maximum likelihood (REML) log likelihood = −2109.4, and the REML AIC = 4286.8.

† The parameter σ^2_{α} is estimated as the variance in the fixed estimates of α_i, and therefore, confidence limits are not computed.

phylogenetic patterns is to be expected, because the occurrence of species among sites in part depends on the second environmental gradient for which species sensitivities have a phylogenetic signal. The performance of model I to identify phylogenetic patterns is similar to the case with a single environmental gradient, with phylogenetic signal found in 87 data sets (Table 4).

Uses of fitted models

PGLMMs give fitted models that can be used to extract information from data sets. To illustrate this, we simulated a data set in which there was a single environmental gradient, phylogenetic signal in species sensitivities to the gradient, and additional site-to-site variation in the number of species (Fig. 6). We then fit these data using model I (Table 5). Because the PGLMMs are binary models, there is no residual variance that is estimated in the statistical model; instead, there is variability generated by the binomial sampling process that cannot be eliminated regardless of how well the model fits the data. The variance estimates of random effects, in this case σ^2_{spp} and σ^2_{site}, give a breakdown of the covariances in species occurrences according to species-specific patterns that depend on phylogeny, σ^2_{spp}, and site-to-site variation in the number of species they contain, σ^2_{site}. Furthermore, the variance σ^2_{α} in the fixed estimates of differences in species mean incidence α_i approximates the contribution of variation in species occurrences across all communities in community structure. For a formal partitioning of variances, it would generally be preferable to treat α_i as a random effect, provided the assumption that α_i follows a Gaussian distribution is valid. The variances occur in logit(μ_i) where $\mu_i = \Pr(Y_i = 1)$, rather than in the probability μ_i itself. Nevertheless, they can still be interpreted as a measure of the proportion of community structure explained by different forces. For this simulated data set, the greater proportion of the variance attributed to σ^2_{spp} (71%; Table 5) in comparison to σ^2_{site} (16%) and σ^2_{α} (13%) indicates that phylogenetically correlated among-species variation in occurrences explains most of the community structure.

In addition to partitioning variances, PGLMMs can also be used to make predictions about the occupancy of sites. To illustrate this, suppose we ask whether species 1 (the leftmost species in Fig. 6) occurs in site 1 (the bottom site in Fig. 6) given that we know something about the presence/absence of phylogenetically related species in site 1. In the absence of additional information, the probability of species 1 occurring in site 1 is 0.38 (Table 6). If species 2 is known to be present, however, the probability that species 1 occurs increases to 0.48, while if species 2 is known to be absent, the probability that species 1 occurs decreases to 0.33. Similarly, the presence or absence of both species 2 and 3 changes the probability for species 1 to 0.54 and 0.29, respectively. As information about more species is obtained, the probability that species 1 occurs in site 1 can be refined (Table 6).

Although we have used only model I for these illustrations, the other PGLMMs can similarly be used to partition variances and make predictions about species co-occurrences. Fitted models can also be used to simulate communities that have the same statistical attributes (to the limit defined by the model) of a real data set. Thus, fitted models become useful tools for exploring properties of data sets.

Table 6. Conditional probability that species 1 occurs in site 1 given the presence or absence of phylogenetically related species (species 2–6) in site 1 (species from left to right in Fig. 6).

Species	Present	Absent
No information	0.38	0.38
2	0.48	0.33
2, 3	0.54	0.29
2, 3, 4	0.59	0.26
2, 3, 4, 5	0.61	0.25
2, 3, 4, 5, 6	0.64	0.23

Notes: The column labeled "Present" gives the probability that species 1 occurs given that the listed species are known to be present, and the column labeled "Absent" gives the cases when the listed species are absent. Probabilities were calculated from the model fit in Table 5.

Discussion

We have demonstrated how PGLMMs can be formulated to address a variety of problems about the phylogenetic structuring of communities while potentially incorporating information about environmental factors and species traits. The PGLMMs provided equal or better statistical power than the alternative methods we tried. Furthermore, the PGLMMs allowed complex hypotheses to be tested in single analyses that use all available data. This compactness simplifies analyses and increases the statistical power of tests.

We also presented two simple examples of the advantages of fitting inferential models to data rather than relying on metrics and randomization tests to identify phylogenetic patterns. PGLMMs are built around hypothesized processes that underlie the statistical distribution of data, and fitting PGLMMs involves estimating parameters that give a match between the observed data and underlying processes that could generate them. Thus, the goal of PGLMMs is not to detect pattern in a single data set, but instead to model the underlying processes that generated the data. Once a PGLMM is fit, it opens up all of the tools of inferential statistics, such as assessing the magnitudes of model parameters, predicting new values of the response variable, and employing diagnostic tests for model goodness of fit (Judge et al. 1985, Gelman and Hill 2007, Bolker et al. 2009). We presented only two simple examples from the wide range of possible questions that could be asked with fitted PGLMMs.

As with any statistical model fitting, care must be taken in formulating an appropriate model. We advocate starting with careful inspection of the data and possibly using established metrics and randomization tests to get an understanding of the data before formulating a PGLMM. Appropriate formulations will depend on both a given data set and the hypothesis of interest. If there are contrasting hypotheses about a given data set that do not involve the magnitude (or difference from zero) of a specific parameter, then model selection (e.g., with Akaike's Information Criterion) can be used to compare different models (Akaike 1973, Burnham and Anderson 2002, Vaida and Blanchard 2005). Because the estimation approach we used is based on restricted maximum likelihood estimation (REML), models should not be compared that differ in fixed effects. This is because REML is based on the conditional likelihood function (Smyth and Verbyla 1996), with resulting estimates of the variance components (random effects) of the model conditioned on the mean components (fixed effects). In practice this might only be a minor limitation, because competing models will most likely differ in random effects. Nonetheless, maximum likelihood (ML) approaches could be used for unrestrained model selection (Appendix B).

The greatest limitation of application of PGLMMs is the absence of general software to perform the necessarily calculations. This will likely change quickly with continued development of versatile packages like lmer in the R statistical language (R Development Core Team 2005, Bates et al. 2008). Our approach (Appendix B) can be used for a wide range of PGLMMs, and simulations show it has good statistical performance (in fact, at least as good as lmer in applications to non-phylogenetic GLMMs; Appendix B). The simulations we presented throughout this paper consisted of 992 points (31 sites and 32 species), and the PGLMMs were fit in a minute or two on an old laptop. For much larger problems (e.g., 10 000 points), however, calculations will be slow. It is also possible to place PGLMMs in a Bayesian estimation framework and use Markov Chain Monte Carlo (MCMC) approaches (Gelman and Hill 2007). Comparing among different estimation approaches, however, is beyond the scope of the present work.

PGLMMs have the flexibility to be constructed to address numerous questions. Although we have focused on presence/absence data, abundances can equally be used. If species abundances are often low, then it is possible to formulate the PGLMMs under the assumption that the response variable (number of individuals of a given species in a site) are Poisson distributed (Gelman and Hill 2007). If present populations are large, then the distribution of the response variable becomes problematic, because it will likely consist of either zeros or large values. In this situation, zero-inflated distributions can be used, although these will require more complex estimation approaches (Gelman and Hill 2007). On the other hand, if there are few zeros (absences) in the data, then the distribution of densities among sites (or some transformation thereof) might be Gaussian, in which case more-simple linear mixed models (LMMs) could be applied. In a phylogenetic context, these would be similar to phylogenetic regressions with phylogenetic signal in the residuals (Gage and Freckleton 2003, Duncan et al. 2007, Lavin et al. 2008, Revell 2010). Because PGLMMs are built on covariance matrices, other sources of data correlations can be included. In particular, spatial correlations that could arise from dispersal limitation or biogeographic history could be included in a covariance matrix (Cressie 1991, Ives and Zhu 2006). Similarly, temporal fluctuations in species abundances generate autocorrelations that can be represented by covariance matrices. Helmus et al. (2010) recently investigated the population dynamics of zooplankton species in several lakes, showing that whole-lake experimental manipulations decreased the phylogenetic diversity of communities. A PGLMM for this type of data could be used to get more information at the species level to explain changes in communities in response to disturbances.

The main message we wish to convey is that fitting models to data is almost always more informative and statistically powerful than the metric/randomization approach that permeates not just analyses of phylogenetic community structure, but also community ecology in general. Recent development of statistical approaches and increased computing power open up possibilities that were unimaginable even 10 years ago. While model-based inferential approaches require users to have a moderate amount of statistical background, GLMMs are becoming commonplace in ecological and evolutionary studies (Bolker et al. 2009), and PGLMMs introduce few challenges beyond those already mastered by many ecologists and evolutionary biologists.

Acknowledgments

Funding for this work was provided by NSF-DEB-0816613 to A. R. Ives. M. R. Helmus was partially funded by a Chinese Academy of Sciences Research Fellowship for International Young Researchers and NSF-DBI-0906011 (Bioinformatics postdoc fellowship).

Literature Cited

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov, and F. Csaki, editors. Second International Symposium on Information Theory. Akadeiai Kiado, Budapest, Hungary.

Bates, D., M. Maechler, and B. Dai. 2008. lme4: linear mixed-effects models using S4 classes. R package version 0.999375-1. R Foundation for Statistical Computing, Vienna, Austria.

Blomberg, S. P., T. Garland, Jr., and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J. S. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution 24:127–135.

Burnham, K. T., and D. R. Anderson. 2002. Model selection and inference: a practical information-theoretic approach. Second edition. Springer, New York, New York, USA.

Cavender-Bares, J., D. D. Ackerly, D. A. Baum, and F. A. Bazzaz. 2004. Phylogenetic overdispersion in Floridian oak communities. American Naturalist 163:823–843.

Cavender-Bares, J., A. Keen, and B. Miles. 2006. Phylogenetic structure of Floridian plant communities depends on taxonomic and spatial scale. Ecology 87(Supplement):S109–S122.

Cavender-Bares, J., K. H. Kozak, P. V. A. Fine, and S. W. Kembel. 2009. The merging of community ecology and phylogenetic biology. Ecology Letters 12:693–715.

Cornwell, W. K., and D. D. Ackerly. 2009. Community assembly and shifts in plant trait distributions across an environmental gradient in coastal California. Ecological Monographs 79:109–126.

Cressie, N. A. C. 1991. Statistics for spatial data. John Wiley and Sons, New York, New York, USA.

Crozier, R. H. 1997. Preserving the information content of species: genetic diversity, phylogeny, and conservation worth. Annual Review of Ecology and Systematics 28:243–268.

Duncan, R. P., D. M. Forsyth, and J. Hone. 2007. Testing the metabolic theory of ecology: allometric scaling exponents in mammals. Ecology 88:324–333.

Faes, C., M. Aerts, H. Geys, L. Bijnens, L. V. Donck, and W. Lammers. 2006. GLMM approach to study the spatial and temporal evolution of spikes in the small intestine. Statistical Modelling 6:300–320.

Faith, D. P. 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation 61:1–10.

Gage, M. J. G., and R. P. Freckleton. 2003. Relative testis size and sperm morphometry across mammals: no evidence for an association between sperm competition and sperm length. Proceedings of the Royal Society of London B 270:625–632.

Gelman, A., and J. Hill. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York, New York, USA.

Graves, G. R., and N. J. Gotelli. 1993. Assembly of avian mix-species flocks in Amazonia. Proceedings of the National Academy of Sciences 90:1388–1391.

Hadfield, J. D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software 33:1–22.

Hardy, O. J., and B. Senterre. 2007. Characterizing the phylogenetic structure of communities by an additive partitioning of phylogenetic diversity. Journal of Ecology 95:493–506.

Helmus, M. R., T. J. Bland, C. K. Williams, and A. R. Ives. 2007*a*. Phylogenetic measures of biodiversity. American Naturalist 169:E68–E83.

Helmus, M. R., W. Keller, M. J. Paterson, N. D. Yan, C. H. Cannon, and J. A. Rusak. 2010. Communities contain closely related species during ecosystem disturbance. Ecology Letters 13:162–174.

Helmus, M. R., K. Savage, M. W. Diebel, J. T. Maxted, and A. R. Ives. 2007*b*. Separating the determinants of phylogenetic community structure. Ecology Letters 10:917–925.

Horner-Devine, M. C., and B. J. M. Bohannan. 2006. Phylogenetic clustering and overdispersion in bacterial communities. Ecology 87:S100–S108.

Ives, A. R., and J. Zhu. 2006. Statistics for correlated data: phylogenies, space, and time. Ecological Applications 16:20–32.

Jabot, F. 2010. A stochastic dispersal-limited trait-based model of community dynamics. Journal of Theoretical Biology 262:650–661.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T.-C. Lee. 1985. The theory and practice of econometrics. Second edition. John Wiley and Sons, New York, New York, USA.

Kembel, S. W., P. D. Cowan, M. R. Helmus, W. K. Cornwell, H. Morlon, D. D. Ackerly, S. P. Blomberg, and C. O. Webb. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464.

Kizilkaya, K., and R. J. Tempelman. 2005. A general approach to mixed effects modeling of residual variances in generalized linear mixed models. Genetics Selection Evolution 37:31–56.

Krackow, S., and E. Tkadlec. 2001. Analysis of brood sex ratios: implications of offspring clustering. Behavioral Ecology and Sociobiology 50:293–301.

Kraft, N. J. B., W. K. Cornwell, C. O. Webb, and D. D. Ackerly. 2007. Trait evolution, community assembly, and the phylogenetic structure of ecological communities. American Naturalist 170:271–283.

Larsen, R. J., and M. L. Marx. 1981. An introduction to mathematical statistics and its applications. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Lavin, S. R., W. H. Karasov, A. R. Ives, K. M. Middleton, and T. Garland, Jr. 2008. Morphometrics of the avian small intestine, compared with non-flying mammals: a phylogenetic approach. Physiological and Biochemical Zoology 81:526–550.

Legendre, P., R. Galzin, and M. L. HarmelinVivien. 1997. Relating behavior to habitat: solutions to the fourth-corner problem. Ecology 78:547–562.

Leibold, M. A., E. P. Economo, and P. Peres-Neto. 2010. Metacommunity phylogenetics: separating the roles of environmental filters and historical biogeography. Ecology Letters 13:1290–1299.

Losos, J. B. 1996. Phylogenetic perspectives on community ecology. Ecology 77:1344–1354.

MathWorks. 2005. MATLAB. MathWorks, Natick, Massachusetts, USA.

May, R. M. 1990. Taxonomy as destiny. Nature 347:129–130.

Mayfield, M. M., M. F. Boni, and D. D. Ackerly. 2009. Traits, habitats, and clades: identifying traits of potential importance to environmental filtering. American Naturalist 174:E1–E22.

McCulloch, C. E., S. R. Searle, and J. M. Neuhaus. 2008. Generalized, linear, and mixed models. John Wiley and Sons, Hoboken, New Jersey, USA.

McGill, B. J., B. A. Maurer, and M. D. Weiser. 2006. Empirical evaluation of neutral theory. Ecology 87:1411–1423.

Milner, J. M., D. A. Elston, and S. D. Albon. 1999. Estimating the contributions of population density and climatic fluctuations to interannual variation in survival of Soay sheep. Journal of Animal Ecology 68:1235–1247.

Peres-Neto, P. R. 2004. Patterns in the co-occurrence of fish species in streams: the role of site suitability, morphology and phylogeny versus species interactions. Oecologia 140:352–360.

Pillar, V. D., and L. D. S. Duarte. 2010. A framework for metacommunity analysis of phylogenetic structure. Ecology Letters 13:587–596.

R Development Core Team. 2005. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Revell, L. J. 2010. Phylogenetic signal and linear regression on species data. Methods in Ecology and Evolution 1:319–329.

Smyth, G. K., and A. P. Verbyla. 1996. A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. Journal of the Royal Statistical Society Series B 58:565–572.

Swenson, N. G. 2009. Phylogenetic resolution and quantifying the phylogenetic diversity and dispersion of communities. PLoS ONE 4:e4390.

Tofts, R., and J. Silvertown. 2000. A phylogenetic approach to community assembly from a local species pool. Proceedings of the Royal Society of London Series B 267:363–369.

Vaida, F., and S. Blanchard. 2005. Conditional Akaike information for mixed-effects models. Biometrika 92:351–370.

Warwick, R. M., and K. R. Clarke. 1998. Taxonomic distinctness and environmental assessment. Journal of Applied Ecology 35:532–543.

Webb, C. O. 2000. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. American Naturalist 156:145–155.

Webb, C. O., D. D. Ackerly, M. A. McPeek, and M. J. Donoghue. 2002. Phylogenies and community ecology. Annual Review of Ecology and Systematics 33:475–505.

Webb, C. O., J. B. Losos, and A. A. Agrawal. 2006. Integrating phylogenies into community ecology. Ecology 87:S1–S2.

Weiblen, G. D., C. O. Webb, V. Novotny, Y. Basset, and S. E. Miller. 2006. Phylogenetic dispersion of host use in a tropical insect herbivore community. Ecology 87:S62–S75.

APPENDIX A

Derivation of $\sigma^2_{\text{repulse}}\mathbf{\Sigma}_{\text{repulse}}$ (*Ecological Archives* M081-018-A1).

APPENDIX B

Statistical estimation and validation of phylogenetic generalized linear mixed models, PGLMMs (*Ecological Archives* M081-018-A2).

SUPPLEMENT

MATLAB computer code for PGLMM estimation and data simulation (*Ecological Archives* M081-018-S1).